REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

AD-A240 678

1

1 is estimated to average 1 hour per response, including the time for reviewing instructions, search-
he data needed, and completing and reviewing the collection of information. Send comments re-
is collection of information, including suggestions for reducing this burden, to Washington Head-
ns and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the

2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
August 1991	Journal Article

4. TITLE AND SUBTITLE
Methods and Design: Measuring Recognition Performance Using Computer-based and Paper-based Methods

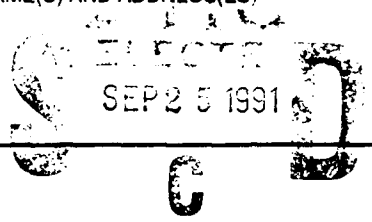
5. FUNDING NUMBERS
None

6. AUTHOR(S)
Pat-Anthony Federico

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Navy Personnel Research and Development Center
San Diego, California 92152-6800

8. PERFORMING ORGANIZATION
JA-91-10

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)



10. SPONSORING/MONITORING
Behavior Research, Methods, Instruments & Computers, 23(3), pp. 341-347, 1991

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT *(Maximum 200 words)*

Using a within-subjects design, we administered to 80 naval pilots and flight officers computer-based and paper-based tests to access recognition of aircraft silhouettes in order to determine the relative reliability and validities of these two measurement modes. Estimates of internal consistencies, equivalences, and discriminative validities were computed for multiple performance measures. It was established that the relative reliabilities and validities derived for these two assessment schemes were contingent on the employed multivariate measurement criteria, that is, percentage correct responses, average response latency, and average degree of confidence in recognition judgments, as well as the statistical criteria used to ascertain the comparative quality of these two modes of testing.

25

91-11466

Computer-based testing, * [illegible]

14. SUBJECT TERMS
aircraft; silhouette; ~~recognition;~~ computer-based testing; measurement; assessment; [illegible]

15. NUMBER OF PAGES
7

16. PRICE CODE

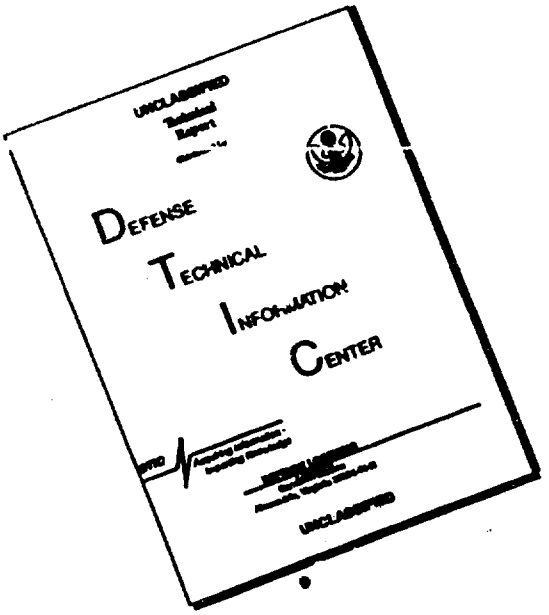
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

NSN 7540-01-280-5500

91 9 25 019

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Behavior Research Methods, Instruments, & Computers
1991, 23 (3), 341-347

— METHODS & DESIGNS —

Measuring recognition performance using computer-based and paper-based methods

PAT-ANTHONY FEDERICO
Navy Personnel Research and Development Center, San Diego, California

Using a within-subjects design, we administered to 83 naval pilots and flight officers computer-based and paper-based tests to assess recognition of aircraft silhouettes in order to determine the relative reliabilities and validities of these two measurement modes. Estimates of internal consistencies, equivalences, and discriminative validities were computed for multiple performance measures. It was established that the relative reliabilities and validities derived for these two assessment schemes were contingent on the employed multivariate measurement criteria, that is, percentage correct responses, average response latency, and average degree of confidence in recognition judgments, as well as the statistical criteria used to ascertain the comparative quality of these two modes of testing.

The consequences of computer-based assessment on examinees' performance are not obvious. The investigations that have been conducted on this topic have produced mixed results. Some studies (D. F. Johnson & Mihal, 1973; Serwer & Stolurow, 1970) demonstrated that test takers do better on verbal items given by computer-based tests than they do on paper-based tests; however, just the opposite was found by other studies (D. F. Johnson & Mihal, 1973; Wildgrube, 1982). One investigation (Sachar & Fletcher, 1978) yielded no significant differences resulting from computer-based and paper-based modes of administration on verbal items. Two studies (English, Reckase, & Patience, 1977; Hoffman & Lundberg, 1976) demonstrated that these two testing modes did not affect performance on memory retrieval items. Sometimes (D. F. Johnson & Mihal, 1973) test takers do better on quantitative tests when they are computer given, sometimes (Lee, Moreno, & Sympson, 1984) they do worse, and other times (Wildgrube, 1982) it may make no difference. Other studies have supported the equivalence of computer-based and paper-based administration (Elwood & Griffin, 1972; Hedl, O'Neil, & Hansen, 1973; Kantor, 1988; Lukin, Dowd, Plake, & Kraft, 1985). Some researchers (Evan & Miller, 1969; Koson, Kitchen, Kochen, & Stodolosky, 1970; Lucas, Mullin, Luna, & McInroy, 1977; Lukin et al., 1985; Skinner & Allen, 1983) have reported psychometric capabilities of computer-based assessment to be comparable or superior to paper-based assessment in clinical settings.

The assistance of Chris Cassella, Glen Little, Don Maffly, Corbin Miller, Dave Setter, and Ellen Schuller is appreciated and acknowledged. Opinions or assertions contained herein are those of the author and are not to be construed as official or reflecting the views of the Department of the Navy. Requests for reprints should be sent to the author at Code 15, Navy Personnel Research and Development Center (NPRDC), San Diego, CA 92152-6800

Investigations of computer-based presentation of personality items have yielded reliability and validity indices comparable to those obtained with typical paper-based presentation (Katz & Dalby, 1981; Lushene, O'Neil, & Dunn, 1974). No significant differences were found in the scores of measures of anxiety, depression, and psychological reactance due to computer-based and paper-based administration (Lukin et al., 1985). Studies of cognitive tests have provided inconsistent findings, with some studies (Hitti, Riffer, & Stuckless, 1971; Rock & Nolen, 1982) demonstrating that the computerized version is a viable alternative to the paper-based version. Other research (Hansen & O'Neil, 1970; Hedl et al., 1973; D. F. Johnson & White, 1980; J. H. Johnson & K. N. Johnson, 1981), though, indicated that interacting with a computer-based system to take an intelligence test could elicit a considerable amount of anxiety, which could affect performance.

Regarding computerized adaptive testing (CAT), some empirical comparisons (McBride, 1980; Sympson, Weiss, & Ree, 1982) yielded essentially no change in validity due to mode of administration. However, test-item difficulty may not be indifferent to manner of presentation for CAT (Green, Bock, Humphreys, Linn, & Reckase, 1984). The effect of switching from paper-based to computer-based administration is thought to have three aspects: (1) an overall mean shift, in which all items may be easier or harder; (2) an item-mode interaction, in which a few items may be altered and others not; and (3) the nature of the task itself, which may be changed by computer administration. A computer simulation study (Divgi, 1988) demonstrated that a CAT version of the Armed Services Vocational Aptitude Battery had a higher reliability than a paper-based version for these subtests: general science, arithmetic reasoning, word knowledge, paragraph comprehension, and mathematics knowledge.

The inconsistent results of mode, manner, or medium of testing may be due to differences in methodology, test content, population tested, or the design of the study (Lee et al., 1984).

With computer costs decreasing and people's knowledge of these systems increasing, it becomes more likely economically and technologically that many benefits can be gained from their use. A direct advantage of computer-based testing is that individuals can respond to items at their own pace, thus producing ideal power tests. Some indirect advantages of computer-based assessment are increased test security, less ambiguity about students' responses, minimal or no paperwork, immediate scoring, and automatic recordkeeping for item analysis (Green, 1983a, 1983b). Some of the strongest support for computer-based assessment is based on the awareness of faster and more economical measurement (Elwood & Griffin, 1972; D. F. Johnson & White, 1980; Space, 1981). Cory (1977) reported some advantages of computerized testing over paper-based testing for predicting job performance.

Ward (1984) stated that computers can be employed to augment what is possible with paper-based measurement (e.g., to obtain more precise information regarding a student than is likely with more customary measurement methods) and to assess additional aspects of performance. He enumerated and discussed potential benefits that may be derived by using computer-based systems to administer traditional tests. Some of these are as follows: (1) individualizing assessment, (2) increasing the flexibility and efficiency of managing test information, (3) enhancing the economic value and manipulation of measurement databases, and (4) improving diagnostic testing. Millman (1984) agreed with Ward that computer-based measurement encourages individualized assessment and that designing software within the context of cognitive science is important. Also, limiting computer-based assessment is not so much hardware inadequacy but incomplete comprehension of the processes intrinsic to testing (Federico, 1980).

Many benefits may be obtained from computerized testing. Some of these may be related to attitudes and assumptions associated with the use of novel media or innovative technology per se. However, and just as readily, potential problems may result from the use of computer-based measurement. Differences between this mode of assessment and traditional testing techniques may or may not affect the reliability and validity of measurement. Notably absent from this literature are studies that have compared the testing characteristics of computer-based assessment with customary measurement methods for assessing recognition performance.

One discernible difference between employing a computer-based method versus a paper-based method for recognition testing of shapes, silhouettes, or spatial forms is the degree of control over stimulus presentation, exposure, or duration. Complete control of stimulus exposure is possible when using a computer-based method, whereas strict manipulation of stimulus presentation is practically impossible or intrinsically lacking when using a paper-based method. One might expect or hypothesize that the longer the stimulus presentation or viewing time of a test item available to the subject during paper-based recognition assessment, the more reliable and valid the measurement compared with computer-based recognition assessment. This assumes that the exposure of each shape or silhouette item during the computer-based test is approaching tachistoscopic durations. If brief exposures of figural forms are employed during computer-based recognition testing, subjects may not have sufficient search time to detect or identify distinctive characteristics, features, or attributes, which are necessary for correct recognition. A salient research issue that should be addressed is the specification of some of the important psychometric implications of employing computer-based versus paper-based procedures for measuring recognition performance. The primary purpose of this research was to evaluate empirically the relative reliability and validity of computer-based and paper-based procedures for assessing recognition of aircraft silhouettes.

METHOD

Subjects

The subjects, who volunteered to participate in this study, were 83 male student pilots and radar intercept officers from the Fleet Replacement Squadron, VF-124, Naval Air Station Miramar. These students must learn to recognize or identify Soviet and non-Soviet aircraft silhouettes so that they can properly employ the F-14 fighter.

Subject Matter

The subject matter consisted of line drawings of the front, side, and top silhouettes of Soviet and non-Soviet aircraft. A paper-based study guide was designed and developed for the subjects to help them learn to recognize the silhouettes of 4 Soviet naval air bombers and 10 of their front-line fighters. Silhouettes of non-Soviet aircraft were also presented, since these could be mistaken for Soviet threats or vice versa.

The subjects were asked to study each Soviet silhouette and its corresponding non-Soviet silhouette and note the distinctive features of each. The correct identification of each Soviet and non-Soviet silhouette according to NATO name and alphanumeric designator (e.g., Foxhound or Mig-31) appeared directly below it. The subjects were told that in the near future, their recognition of these Soviet and non Soviet aircraft would be assessed via computer and traditional testing.

In addition to using the paper-based study guide, the subjects were required to learn the silhouettes via the computer-based system described below, which was configured in a training mode for this purpose. In this mode, when a student pressed a key, a silhouette would reappear together with its correct identification so that they could be associated.

Computer-Based Assessment

Computerized line drawings were used to assess how well the subjects recognized or identified the silhouettes.

These were digitized facsimiles of those employed in the paper-based study guide. A computer game based on a sequential recognition paradigm was developed (Little, Maffly, Miller, Setter, & Federico, 1985). It randomly selected and presented on a computer display, at an arbitrary exposure setting, the front, side, or top views of 4 Russian bombers and 10 of their advanced fighters. For this research, the exposure of a silhouette on the computer screen was approximately 500 msec. Also, the game management system can choose and flash corresponding silhouettes of NATO aircraft, which act as distractors because of their high degree of similarity to the Soviet silhouettes. The subjects' task was to identify as quickly as possible the aircraft that was represented by each silhouette. The subjects entered on the keyboard what they recognized each aircraft to be, using its NATO name or corresponding alphanumeric designation. Misspellings counted as wrong responses.

This particular computer-based game or test assesses student performance by measuring the "hit rate," or number of correct recognitions, out of a total of 42 silhouettes, half of which are Soviet and the other half non-Soviet; the time, or latency, it takes a student to make a recognition judgment for each target or distractor aircraft; and the degree of confidence the student has in each of his recognition decisions. At the end of the game, feedback is given to the student in terms of his hit rate (computer-based total percentage correct responses, or CTP), average response latency (computer-based total average response latency, or CTL), average degree of confidence in his recognition judgments (computer-based total average degree of confidence, or CTC), and how his performance compares to other students who have played the game.

Paper-Based Assessment

Since the computer-based test randomly selected and presented silhouettes, creating a distinct sequence of test items or form for each subject, an attempt was made to simulate this computer-based administration of different forms by employing different paper-based forms. Consequently, two alternative forms of a paper-based test were designed and developed to assess the subjects' recognition of the silhouettes mentioned above. The alternative test forms mimicked as much as possible the format used by the computer-based test. Also, these paper-based alternative forms employed as individual items facsimiles of the digitized silhouettes used in their computer counterparts. Both test forms were presented as booklets, each containing 42 items representing the front, top, or side silhouettes of aircraft. The subjects' task was to identify as quickly as possible the aircraft that was represented by each item's silhouette. They were asked to write in the space provided what they recognized the aircraft to be, using its NATO name or corresponding alphanumeric designation. Misspellings counted as wrong responses. The subjects were instructed not to turn back to previous pages in the test booklet to complete items they had left blank. The students were encouraged to go through the test items quickly to approximate as much as possible the silhouette exposure employed by the computer-based test. The subjects were closely monitored to assure that they complied with this procedure.

After the subjects wrote down what they thought an aircraft was, they were required to indicate on a scale that appeared below each silhouette the degree of confidence in their recognition decision concerning the specific item. Like the confidence scale used for the computer-based test, this one went from least confident, or 0% confidence, in their recognition decision on the left, to most confident, or 100% confidence, on the right, using a 10-point scale. The subjects were instructed to use this confidence scale by placing a check mark at the point that best reflected or approximated the sureness they had in their judgment. To learn how to respond properly to the silhouette test items, the subjects were asked to look at three completed examples. A subject's percentage of correct recognitions (paper-based total percentage correct responses, or PTP) and average degree of confidence (paper-based total average degree of confidence, or PTC) for the paper-based test were measured and recorded.

Procedure

Prior to testing, the subjects learned to recognize the aircraft silhouettes using two media: (1) a paper-based form structured as a study guide, and (2) a computer-based form using the system in the training mode, as mentioned above. Mode of assessment, computer-based or paper-based, was manipulated as a within-subjects variable (Kirk, 1968). All subjects were administered the paper-based test before the computer-based test. The two forms of the paper-based tests were counterbalanced or alternated in their administration to the subjects. After the subjects received the paper-based test, they were immediately administered the computer-based test. It was assumed that a subject's state of recognition knowledge was the same during the administration of both tests. The subjects took approximately 10–15 min to complete the paper-based test and 15–20 min to complete the computer-based test. This difference in completion time was primarily due to lack of typing proficiency among some of the subjects.

Reliabilities for both modes of testing were estimated by deriving internal consistency indices using an odd–even item split. These reliability estimates were adjusted by employing the Spearman-Brown Prophecy Formula (Thorndike, 1982). Reliability estimates were calculated for test score, average degree of confidence, and average response latency for the computer-based test; reliability estimates were calculated only for test score and average degree of confidence for the paper-based test. Estimates were not computed for average response latency since this was not measured for the paper-based test. Equivalences between these two modes of assessment were estimated by Pearson product-moment correlations for total test score and average degree of confidence

To derive discriminative validity estimates, we placed the research subjects into two groups according to whether

or not their performance through the squadron's curriculum was above or below the mean grade for this sample. A stepwise discriminant analysis (Dillon & Goldstein, 1984), using Wilks's criterion for including and rejecting variables, and associated statistics were computed to ascertain how well computer-based and paper-based measures distinguished among the defined groups expected to differ in their recognition of aircraft silhouettes.

RESULTS

Reliability and Equivalence Estimates

The means, standard deviations, intercorrelations, and associated statistics for the computer-based and paper-based measures of recognition performance are presented in Table 1. As can be seen, the subjects' recognition performance was significantly better, and they had significantly more confidence, on the paper-based test than on the computer-based test. On both the computer-based and paper-based versions, the subjects' recognition performance correlated significantly and positively with confidence in their identification judgments. For the computer-based mode, the subjects' response latency varied inversely and significantly with their recognition performance and confidence. It appears that recognition performance and confidence were more strongly associated for the paper-based test than for its computer-based counterpart.

Split-half reliability and equivalence estimates of computer-based and paper-based measures of recognition performance are presented in Table 2. The adjusted reliability estimates are relatively high, ranging from .89 to .97. The difference in reliabilities for computer-based and paper-based measures for average degree of confidence was statistically significant ($p < .02$), using a test described by Edwards (1964, p. 85). However, the difference in reliabilities for computer-based and paper-based measures of the recognition test score was not significant. These results revealed that (1) the computer-based and paper-based measures of test score were not significantly different in reliability or internal consistency, and (2) the paper-based measure of average degree of confidence was more reliable or internally consistent than the computer-based measure.

Estimates of equivalence between corresponding computer-based and paper-based measures of recognition test score and average degree of confidence were .67 and .81, respectively. These suggested that the computer-based and paper-based measures had from 45% to 66% variance in common, implying that these different modes of assessment were only partially equivalent. The equivalences for test score and average degree of confidence measures were significantly different ($p < .001$). This result suggested that computer-based and paper-based measures of average degree of confidence were more equivalent than the measures of recognition test score.

Table 1
Means, Standard Deviations, and Intercorrelations of Computer-Based and Paper-Based Measures of Recognition Performance

	M	*SD*	CTP	CTC	CTL	PTP
CTP	77.39†	20.59				
CTC	89.93†	12.89	.57			
CTL	1779.51	1493.56	−.22*	−.45		
PTP	83.81†	17.05	.67	.72	−.33	
PTC	92.14†	10.38	.48	.81	−.41	.69

Note—CTP = computer-based total percentage correct responses; CTL = computer-based total average response latency; CTC = computer-based total average degree of confidence; PTP = paper-based total percentage correct responses; PTC = paper-based total average degree of confidence. CTL was measured in milliseconds. $r(81) > .27$, $p < .005$. *$r(81) > .21$, $p < .025$. †$t(82) = -3.77$, $p < .001$. ‡$t(82) = -2.70$, $p = .008$.

Table 2
Split-Half Reliability and Equivalence Estimates of Computer-Based and Paper-Based Measures of Recognition Performance

	Reliability		
Measure	Computer-Based	Paper-Based	Equivalence
Score	.90	.89	.67
Confidence	.95	.97	.81
Latency	.93		

Note—Split-half reliability estimates were adjusted by using the Spearman-Brown Prophecy Formula

Discriminative Validity

The discriminant analysis was computed to determine how well computer-based and paper-based measures of recognition performance differentiated groups defined by above or below mean average curriculum grade. The statistics associated with the single significant function, standardized discriminant-function coefficients, pooled within-groups correlations between the discriminant function and computer-based and paper-based measures, and group centroids for above or below mean average curriculum grade are presented in Table 3. The discriminant-function coefficients, which consider the interrelationships or interdependencies among the multivariate measures, revealed the relative contribution or comparative importance of the variables in defining this derived dimension to be CTC, PTC, PTP, CTP, and CTL, respectively. The within-groups correlations, which were computed for each individual measure partialling out the interrelationships of all the other variables, indicated that the major contributors to the discriminant function were CTP, CTC, and CTL, respectively, all computer-based measures.

The means and standard deviations for groups above or below mean average curriculum grade, univariate *F* ratios, and levels of significance for computer-based and paper-based measures of recognition performance are summarized in Table 4. Considering the measures as univariate variables (i.e., independent of their multivariate relationships or dependencies with one another), these statistics revealed that one computer-based measure, CTL, and one paper-based measure, PTC, significantly differentiated the two groups. The means revealed that the group above mean curriculum grade had shorter computer-based latencies than the group below mean curriculum grade,

Table 3
Statistics Associated with the Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-Based Measures, and Group Centroids for Above or Below Mean Curriculum Grade

Discriminant Function					
Eigen Value	Canonical Correlation	Wilks's Lambda	Chi-Square	*df*	*p*
.14	.35	.88	9.98	5	.076

Measure	Discriminant Coefficient	Within-Group Correlation	Group	Centroid
CTP	−.60	.60	Above Mean Grade	.32
CTC	−.97	−.55	Below Mean Grade	−.42
CTL	−.52	.48		
PTP	.80	.25		
PTC	94	−.03		

Table 4
Means and Standard Deviations for Groups Above or Below Mean Grade, Univariate *F* Ratios, and Levels of Significance for Computer-Based and Paper-Based Measures

		Group			
Measure		Above Mean Grade (n = 47)	Below Mean Grade (n = 36)	*F*	*p*
CTP	*M*	77.19	77.64	.01	.92
	SD	18.48	23.33		
CTC	*M*	90.99	88.54	.73	.39
	SD	12.74	13.06		
CTL	*M*	1,522.06	2,115.61	3.31	.07
	SD	1,554.12	1,359.19		
PTP	*M*	86.40	80.42	2.56	.11
	SD	16.65	17.19		
PTC	*M*	94.09	89.61	3.92	.05
	SD	9.47	11.09		

and that the former group had a higher paper-based average degree of confidence than the latter group.

The discriminant-function coefficients and group means also implied that the students with above mean average grades (1) did relatively well on the paper-based test (PTP) and relatively poorly on the computer-based test (CTP), and (2) had more confidence in their paper-based performance (PTC) than their computer-based performance (CTC). These statistics together with the discriminant-function coefficients and group means reported for CTL and CTP as well as the correlations between CTL and CTP and CTC suggested that there may have been a slight speed–accuracy tradeoff for computer-based recognition testing.

In general, the multivariate and subsequent univariate results established that according to two sets of criteria, the discriminant coefficients and *F* ratios and corresponding means, the discriminant validities of computer-based and paper-based measures were about the same for distinguishing groups above or below mean average curriculum grade. However, according to another set of criteria, the pooled within-groups correlations between the discriminant function and the computer-based and paper-based measures, the former had superior discriminative validity to the latter.

DISCUSSION

This study established that the relative reliability of computer-based and paper-based measures depends on the specific criterion assessed. That is, regarding the recognition test score itself, it was found that computer-based and paper-based measures were not significantly different in reliability or internal consistency. However, regarding the average degree of confidence in recognition judgments, it was found that the paper-based measure was more reliable or internally consistent than its computer-based counterpart. The extent of the equivalence between these two modes of measurement was contingent on particular performance criteria. It was demonstrated that the equivalence of computer-based and paper-based measures of average degree of confidence was greater than that for recognition test score. The relative discriminative validity of computer-based and paper-based measures was dependent on the specific statistical criteria selected. The discriminant coefficients, *F* ratios, and corresponding means indicated that the validities of computer-based and paper-based measures were about the same for distinguishing groups above or below mean curriculum grade. However, according to another set of criteria, the pooled within-groups correlations between the discriminant function and computer-based and paper-based measures, the former had better validity than the latter.

Even though the subjects had more time to view each silhouette during the paper-based test than during the computer-based test, recognition scores for the two measurement modes were not significantly different in reliability. However, the longer exposures of paper-based assessment seemed to have improved the reliability of measuring the subjects' degree of confidence in their recognition judgments compared with computer-based assessment. As hypothesized, the longer exposures intrinsic to the paper-based method seemed to have facilitated the subjects' recognition scores. They performed significantly better on the paper-based test than on the computer-based test. Also, it appears that the difference in the silhouette exposures of the two testing methods had greater impact on the equivalence of recognition test score than on the average degree of confidence. This seemed so, since the equivalence of the computer-based and paper-based measures of recognition performance was significantly less than the degree of confidence. The inherent difference in silhouette viewing times between the computer-based and paper-based assessment modes was expected to affect the recognition process itself more than the degree of confidence in identification judgment.

The results of this research supported the findings of some studies, but not others. Federico and Liggett (1988,

1989) administered computer-based and paper-based tests of semantic knowledge (Liggett & Federico, 1986) to determine the relative reliability and validity of these two modes of assessment. Estimates of internal consistencies, equivalences, and discriminant validities were computed. They established that computer-based and paper-based measures (i.e., test score and average degree of confidence) were not significantly different in reliability or internal consistency. This finding partially agrees with the corresponding result of the present study, since computer-based and paper-based measures of test score were found to be equally reliable; however, the computer-based measure of average degree of confidence was found to be less reliable than its paper-based counterpart. A few of the Federico and Liggett findings were ambivalent, since some results suggested that equivalence estimates for computer-based and paper-based measures (i.e., test score and average degree of confidence) were about the same, and another suggested that these estimates were different. Some of their reported results are different from those established in the present study, in which computer-based and paper-based measures of test score were less equivalent than these measures of average degree of confidence. Finally, Federico and Liggett demonstrated that the discriminative validity of the computer-based measures was superior to that of the paper-based measures. This result is in partial agreement with that found in the present research, where it was also established with respect to some statistical criteria. However, according to other criteria, the discriminative validity of computer-based and that of paper-based measures were about the same.

The results of our present research supported some studies, but not others. Hofer and Green (1985) were concerned that computer-based assessment would introduce irrelevant or extraneous factors that would likely degrade test performance. These computer-correlated factors may alter the nature of the task to such a degree that it would be difficult for a computer-based test and its paper-based counterpart to measure the same construct or content. This could affect reliability, validity, and normative data, as well as other assessment attributes. Several plausible reasons, according to Hofer and Green, may contribute to different performances on these distinct kinds of testing: (1) state of anxiety instigated when confronted by computer-based testing, (2) lack of computer familiarity on the part of the test taker, and (3) changes in response format required by the two modes of assessment. These different dimensions could result in tests that are nonequivalent; however, in our present research these diverse factors had no apparent impact.

On the other hand, a number of known differences between computer-based and paper-based assessment that may affect equivalence and validity (Green, 1986): (1) Passive omitting of items is usually not permitted on computer-based tests. An individual must respond in a different manner than on paper-based tests. (2) Computerized tests typically do not permit backtracking. The test taker cannot easily review items, alter responses, or delay attempting to answer questions. (3) The capacity of the computer screen can have an impact on what usually are long test items, for example, paragraph comprehension. These may be shortened to accommodate the computer display, thus partially changing the nature of the task. (4) The quality of computer graphics may affect the comprehension and degree of difficulty of the item. (5) Pressing a key or using a mouse may be easier than marking an answer sheet. This may affect the validity of speeded tests. (6) Since the computer typically displays items individually, traditional time limits are no longer necessary.

Assuming that these abstract distinctions may affect the equivalence and validity of computer-based and paper-based assessment, the omission of items and backtracking on paper-based tests in this research were not permitted in order to simulate computer-based tests. Computer screen capacity was of no consequence in this study since none of the test items were long. The graphics used in the paper-based recognition test were screen dumps of the actual aircraft silhouettes used in its computer-based counterpart. That is, these images were of the same quality. In this study, neither the computer-based or paper-based measurement employed true speeded tests. Also, we attempted to mimic the individual display of items on the computer-based tests used in this research by closely monitoring the subjects as they took the paper-based test and reminding them to expedite responding without retracing.

When evaluating or comparing different media for instruction and assessment, the newer medium may simply be perceived as being more interesting, engaging, and challenging by the students. This novelty effect seems to disappear as rapidly as it appears. However, in research studies conducted over a relatively short time span, for example, a few days or months at the most, this effect may still linger and affect the evaluation by enhancing the impact of the more novel medium (Colvin & Clark, 1984); this effect could have occurred in our present research. When matching media to distinct subject matters, course contents, or core concepts, some research evidence (Jamison, Suppes, & Welles, 1974) indicates that, other than in obvious cases, just about any medium will be effective for different content. Extrapolating this notion to the measurement domain, the validity results of this study seem to suggest, contrary to the above, that different media may be differentially effective in testing the same subject matter.

REFERENCES

COLVIN, C., & CLARK, R. E. (1984). Instructional media vs. instructional methods. *Performance & Instruction Journal*, **23**, 1-3.

CORY, C. H. (1977). Relative utility of computerized versus paper-and-pencil tests for predicting job performance. *Applied Psychological Measurement*, **1**, 551-564.

DILLON, W. R., & GOLDSTEIN, M. (1984). *Multivariate analysis: Methods and applications*. New York: Wiley.

DIVGI, D. R. (1988). *Two consequences of improving a test battery* (CRM 88-171). Alexandria, VA: Center for Naval Analyses.

EDWARDS, A. L. (1964). *Experimental design in psychological research*. New York: Holt, Rinehart & Winston.

ELWOOD, D. L., & GRIFFIN, R. H. (1972). Individual intelligence testing without the examiner: Reliability of an automated method. *Journal of Consulting & Clinical Psychology*, **38**, 9-14.

ENGLISH, R. A., RECKASE, M. D., & PATIENCE, W. M. (1977). Application of tailored testing to achievement measurement. *Behavior Research Methods & Instrumentation*, **9**, 158-161.

EVAN, W. M., & MILLER, J. R. (1969). Differential effects of response bias of computer versus conventional administration of a social science questionnaire. *Behavioral Science*, **14**, 216-227.

FEDERICO, P-A. (1980). Adaptive instruction: Trends and issues. In R. E. Snow, P-A. Federico, & W. E. Montague (Eds.), *Aptitude, learning, and instruction: Vol. 1. Cognitive process analyses of aptitude* (pp. 1-26). Hillsdale, NJ: Erlbaum.

FEDERICO, P-A., & LIGGETT, N. L. (1988, April). *Comparing computer-based and paper-based assessment strategies for semantic knowledge*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

FEDERICO, P-A., & LIGGETT, N. L. (1989). *Computer-based and paper-based measurement of semantic knowledge* (NPRDC TR 89-4). San Diego, CA: Navy Personnel Research and Development Center.

GREEN, B. F. (1983a). Adaptive testing by computer. *Measurement, Technology, & Individuality in Education*, **17**, 5-12.

GREEN, B. F. (1983b). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A Festschrift in honor of Frederic Lord* (pp. 69-80). Hillsdale NJ: Erlbaum.

GREEN, B. F. (1986, May). *Construct validity of computer-based tests*. Paper presented at the Test Validity Conference, Educational Testing Service, Princeton, NJ.

GREEN, B. F., BOCK, R. D., HUMPHREYS, L. G., LINN, R. L., & RECKASE, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, **21**, 347-360.

HANSEN, D. H., & O'NEIL, H. F. (1970). Empirical investigations versus anecdotal observations concerning anxiety and computer-assisted instruction. *Journal of School Psychology*, **8**, 315-316.

HEDL, J. J., O'NEIL, H. F., & HANSEN, D. H. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting & Clinical Psychology*, **40**, 217-222.

HITTI, F. J., RIFFER, R. L., & STUCKLESS, E. R. (1971). *Computer-managed testing: A feasibility study with deaf students*. Rochester, NY: National Technical Institute for the Deaf.

HOFER, P. J., & GREEN, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting & Clinical Psychology*, **53**, 826-838.

HOFFMAN, K. I., & LUNDBERG, G. D (1976). A comparison of computer-monitored group tests with paper-and-pencil tests. *Educational & Psychological Measurement*, **36**, 791-809.

JAMISON, D., SUPPES, P., & WELLES, S. (1974). The effectiveness of alternative media: A survey. *Annual Review of Educational Research*, **44**, 1-68.

JOHNSON, J. H., & JOHNSON, K. N. (1981). Psychological considerations related to the development of computerized testing stations. *Behavior Research Methods & Instrumentation*, **13**, 421-424.

JOHNSON, D. F., & MIHAL, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist*, **28**, 694-699.

JOHNSON, D. F., & WHITE, C. B. (1980). Effects of training on computerized test performance in the elderly. *Journal of Applied Psychology*, **65**, 357-358.

KANTOR, J. (1988). *The effects of anonymity, item sensitivity, trust, and method of administration on response bias on the job description index*. Unpublished doctoral dissertation, California Schoool of Professional Psychology, San Diego, CA.

KATZ, L., & DALBY, J. T. (1981). Computer-assisted and traditional psychological assessment of elementary-school-age children. *Contemporary Educational Psychology*, **6**, 314-322.

KIRK, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Brooks/Cole.

KOSON, D., KITCHEN, C., KOCHEN, M., & STODOLOSKY, D. (1970). Psychological testing by computer: Effect on response bias. *Educational & Psychological Measurement*, **30**, 808-810.

LEE, J. A., MORENO, K. E., & SYMPSON, J. B. (1984, April). *The effects of mode of test administration on test performance*. Paper presented at the annual meeting of the Eastern Psychological Association, Baltimore, MD.

LIGGETT, N. L., & FEDERICO, P-A. (1980). *Computer-based system for assessing semantic knowledge: Enhancements* (NPRDC TN 87-4). San Diego, CA: Navy Personnel Research and Development Center.

LITTLE, G. A., MAFFLY, D. H., MILLER, C. L., SETTER, D. A., & FEDERICO, P-A. (1985). *A computer-based gaming system for assessing recognition performance (recog)* (TL 85-3). San Diego, CA: Training Laboratory, Navy Personnel Research and Development Center.

LUCAS, R. W., MULLIN, P. J., LUNA, C. D., & MCINROY, D. C. (1977). Psychiatrists and a computer as interrogators of patients with alcohol related illnesses: A comparison. *British Journal of Psychiatry*, **131**, 160-167.

LUKIN, M. E., DOWD, E. T., PLAKE, B. S., & KRAFT, R. G. (1985). Comparing computerized versus traditional psychological assessment. *Computers in Human Behavior*, **1**, 49-58.

LUSHENE, R. E., O'NEIL, H. F., & DUNN, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment*, **34**, 353-361.

MCBRIDE, J. R. (1980). Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 4-15). Minneapolis: University of Minnesota, Department of Psychology.

MILLMAN, J. (1984). Using microcomputers to administer tests: An alternate point of view. *Educational Measurement: Issues & Practices*, **3**, 20-21.

ROCK, D. L., & NOLEN, P. A. (1982). Comparison of the standard and computerized versions of the Raven Coloured Progressive Matrices test. *Perceptual & Motor Skills*, **54**, 40-42.

SACHAR, J. D., & FLETCHER, J. D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 403-419). Minneapolis: University of Minnesota, Department of Psychology.

SERWER, B. L., & STOLUROW, L. M. (1970). Computer-assisted learning in language arts. *Elementary English*, **47**, 641-650.

SKINNER, H. A., & ALLEN, B. A. (1983). Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. *Journal of Consulting & Clinical Psychology*, **51**, 267-275.

SPACE, L. G. (1981). The computer as psychometrician. *Behavior Research Methods & Instrumentation*, **13**, 595-606.

SYMPSON, J. B., WEISS, D. J., & REE, M. (1982). *Predictive validity of conventional and adaptive tests in an air force training environment* (AFHRL-TR-81-40). San Antonio, TX: Brooks Air Force Base, Air Force Human Resources Laboratory.

THORNDIKE, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

WARD, W. C. (1984). Using microcomputers to administer tests. *Educational Measurement: Issues & Practices*, **3**, 16-20.

WILDGRUBE, W. (1982, July). *Computerized testing in the German Federal Armed Forces—Empirical approaches*. Paper presented at the 1982 Computerized Adaptive Testing Conference, Spring Hill, MN.

(Manuscript received September 24, 1990;
revision accepted for publication February 11, 1991.)